오피니언 마이닝과 협업필터링을 이용한 도서 추천시스템

*윤원탁, *박두순
*순천향대학교 컴퓨터소프트웨어공학과
e-mail : ekdma7379@naver.com

Book recommendation system using collaborative filtering and opinion mining

*Won-Tak Yoon,, *Doo-Soon Park+
*Dept. of Computer Software Engineering, SoonChunHyang University

요 약

빅데이터가 일상이 된 현대 사회에서 책 시장의 증가와 책 양의 증가로 인하여 책을 개인에 맞게 선택하는데 어려움이 있다. 그래서 개인 맞춤 추천 시스템이 필요하다. 개인 맞춤 추천 시스템에서 가장 많이 사용하는 방법은 협업 필터링 방법이 있다. 협업 필터링은 희박성 문제를 가지고 있다.

본 논문에서는 협업 필터링 방법에 희박성 문제를 해결하기 위하여 지역, 나이, 성별, 장르 등 개인 성향을 이용하고, 기존의 책 리뷰를 오피니언 마이닝 기법을 적용하여 개인 맞춤형 도서를 추천하는 추천시스템을 제안한다.

1.서론

추천 시스템은 1990년대 협업 필터링에 대한 연구가 시작된 이후 지속적으로 발전해왔다. 특히 전자상거래의 규모가 커지고 있는 만큼 기업에서 추천 시스템을 도입하여 성공적인 효과를 본 사례도 많아지고 있다. 아마존은 1996년 책 추천 시스템을 도입한 이후 사용자의 피드백을 반영하여 성능을 발전시켜왔고, 2015년에는 시스템에 의한 페이지 유입이 전체 유입량의 30%에 달할 만큼 효과적인 것으로 나타났다[1]. 또한, 아마존의 킨들 파이어, 구글의 넥서스 7 등 가격이 저렴한 태블릿PC들이 등장하고 전자책 등 디지털 콘텐츠가 늘어날수록 E러닝 시장은 급격히 성장하고 있으며 중국·인도 등 많은 인구와 높은 성장률, 교육열을 지닌 개발도상국들이 E러닝 시장에 점차 진입하는 것 역시 호재다[2]. 세계 전자책 시장의 성장률은 (그림 1)과 같다.

이처럼 책 시장이 성장하고 데이터의 양이 증가함에 따라 사람들은 책을 선택하는데 많은 어려움을 느끼고 있다. 이러한 문제점을 해결하기 위해 많은 알고리즘과 연구가 진행되고 있다.

(그림 1) 종편 4사 월간 평균 시청률[1]

본 논문에서는 협업 필터링 방법에 희박성 문제를 해결하기 위하여 지역, 나이, 성별, 장르 등 개인 성향을 이용하고, 기존의 책 리뷰를 오피니언 마이닝 기법을 적용하여 R을 사용하여 개인 맞춤형 도서를 추천하는 추천시스템을 제안한다.

2. 도서 추천 시스템의 구성

협업 필터링(Collaborative Filtering)이란 많은 사용자들로부터 얻은 기호정보(taste information)에 따라 사용자들의 관심사들을 자동적으로 예측하게 해주는 방법이다[3]. 협업 필터링 접근법의 근본적인 가정은 사용자들의 과거의 경향이 미래에서도 그대로 유지될 것이라는 전제

+ 본 연구는 NRF-2017R1A2B1008421에 의해 지원되었음

에 있다[4]. 이러한 협업 필터링은 일반적으로 사용자 기반 시스템과 제품 기반 추천 시스템으로 구분할 수 있다.

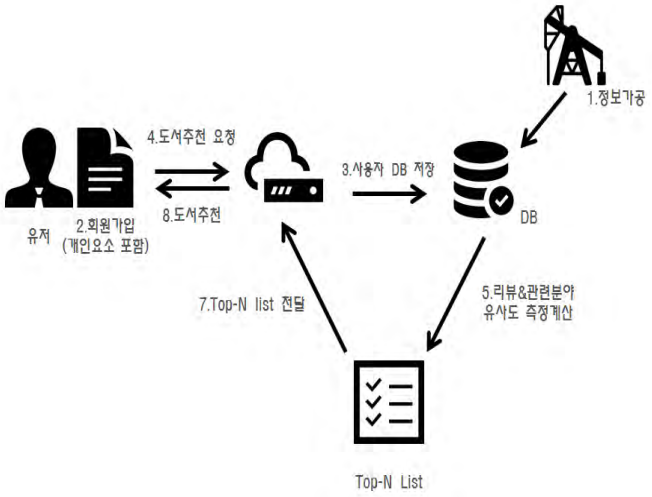
사용자의 유사도를 측정하는 방법에는 여러 가지가 있지만 가장 널리 알려진 방법 중 하나인 피어슨 상관계수를 사용했으며 피어슨 상관계수는 (그림 3)과 같다.

$$s_{ij} = \frac{\sum_u (R_{ui} - \overline{R_i}) \times (R_{uj} - \overline{R_j})}{\sqrt{\sum_u (R_{ui} - \overline{R_i})^2 \times \sum_u (R_{uj} - \overline{R_j})^2}}, -1 \le s_{ij} \le 1$$

(그림 3)피어슨 상관계수[5]

(그림 3)에서 r_{ij}는 사용자 i, j간의 유사도이다. 또한 S_{ik}는 고객 i가 도서 k에 대해 평가한 선호도이고, $\overline{S_i}$는 고객 i의 선호도의 평균이다. S_{jk}는 고객 j가 도서 k에 대해 평가한 선호도이고, $\overline{S_j}$는 고객 j의 선호도의 평균이다[5].

(그림 2)은 논문에서 구현한 시스템의 시나리오이다.



(그림 2) 도서 추천 시스템 구성도

(그림 2)의 시나리오를 자세하게 설명하면 다음과 같다.

1. yes24 홈페이지의 책 분야에서 책에 대한 정보를 저장한다.

2. 지역, 나이, 성별, 장르 등 개인 성향 등 사용자의 유사도를 측정하기 위해 필요한 개인 성향을 입력한다.

3. 회원가입이 된 회원의 요소를 데이터베이스에 저장한다.

4. 회원가입이 된 회원이 도서추천을 요구한다.

5. 피어슨 상관계수를 이용하여 지역, 나이, 성별, 장르 등 개인 성향 등에 대한 유사도를 계산한다. 또한 추가적으로 추천의 정확성을 증가시키기 위해 책의 리뷰 점수와 리뷰에서의 긍정어, 부정어 단어 카운팅, 관심 분야 관련 단어 카운팅을 이용하여 계산한다.

<표 1>은 긍정어, 부정어 단어 카운팅에 쓰일 감성어를 분류해놓은 것이다.

<표 1> 도서 관련 감성어 분류[6]

긍정어	부정어
슥슥, 술술, 재미, 추천, 위로, 흥미, 감동, 감탄, 신선, 훌륭, 따뜻, 인생, 힐링, 선호, 여운, 관심, 권함, 흡입력, 매력, 유쾌	진부, 혼란, 별로, 억지, 아쉬움, 무거움, 실망, 불쾌, 충격, 지루, 불편, 피폐, 난해, 허술, 피곤, 부정, 비추, 답답, 저급, 허탈

리뷰에서 얻을 수 있는 점수는 긍정의 단어가 나올 때마다 +0.01점을, 부정의 단어가 나올 때 마다 -0.01점으로 계산한다. 또한 관심분야의 단어가 나올 때 마다 +0.01점으로 계산한다. 관심분야는 정보를 가지고 온 yes24가 분류한 책 분야에 따라서 분류다. 관심분야에서 최대의 점수는 -0.2~0.2까지이다. 이를 식으로 표현하면 (그림 4)와 같다.

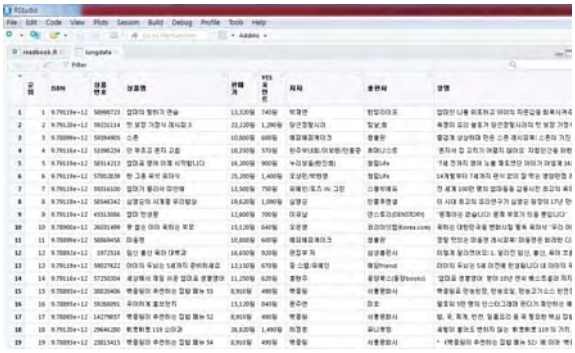
$$S_{ij} + \frac{\text{감성어, 관심분야 점수}}{2.0}$$

(그림 4) 최종 추천 점수

3. 도서 추천 시스템의 구현

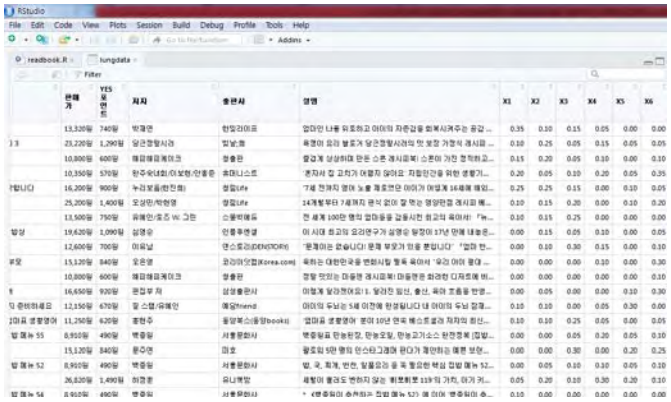
회원가입이 완료가 되면 회원들의 정보를 데이터베이스에 저장한다.

책 정보로는 ISBN, 상품번호, 상품명, 판매가, 저자, 설명, 평점등이 포함돼 있다. 이를 사용하기 위해서 R에서 제공하는 read.csv 함수를 이용하여 책의 정보를 저장한다. (그림 7)은 읽어드린 책의 정보를 R studio에서 출력한 화면이다.



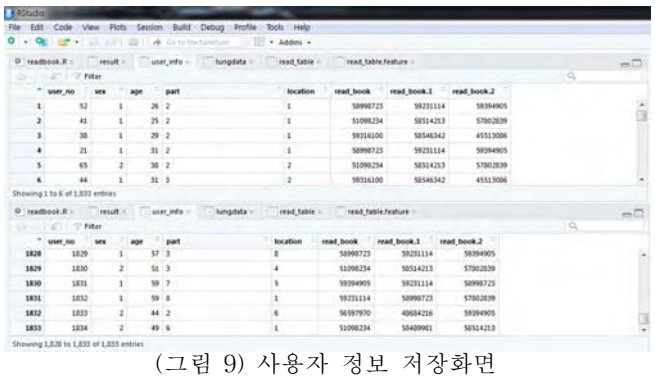
(그림 7)R studio로 읽어드린 책 정보

저장된 책 속에서 사용자가 선택한 관심분야의 관련 단어를 카운팅하기 위해서는 설명의 내용에 명사를 추출한다. 명사를 추출하기 위해서 R Studio에서 사용할 수 있는 한글 자연어 분석 패키지인 KoNLP(Korean Natural Language Processing) 패키지를 사용한다. 추출한 명사에 대하여 관련분야의 단어를 카운팅하고 이를 다시 저장한다. (그림 8)는 관련분야의 단어 점수를 저장한 화면이다.



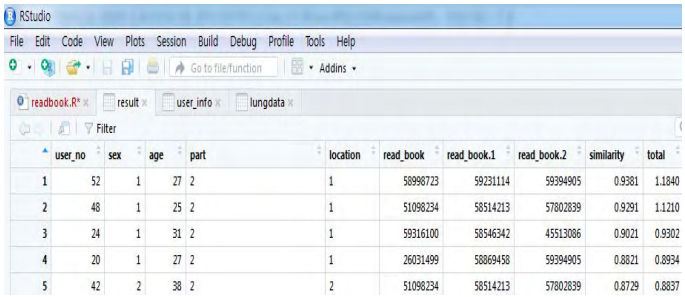
(그림 8) 관련분야 단어 카운팅한 점수를 저장한 테이블

(그림 9)는 1834명의 사용자 정보가 저장되어있는 DB의 정보를 가져온 화면이다. 과거에 읽은 책의 정보는 yes24의 상품번호로 저장되어 있다.



(그림 9) 사용자 정보 저장화면

(그림 10)은 이 중에서 서울에 사는 관심분야를 문학으로 설정한 26살 남자 사용자의 유사도 측정결과이다.



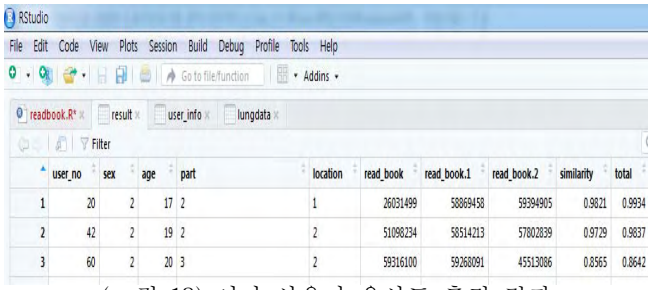
user_no	sex	age	part	location	read_book	read_book.1	read_book.2	similarity	total
1	52	1	27	2	1	58998723	59231114	59394905	0.9381	1.1840
2	48	1	25	2	1	51098234	58514213	57802839	0.9291	1.1210
3	24	1	31	2	1	59316100	58546342	45513086	0.9021	0.9302
4	20	1	27	2	1	26031499	58869458	59394905	0.8821	0.8934
5	42	2	38	2	2	51098234	58514213	57802839	0.8729	0.8837

(그림 10) 남자 사용자 유사도 측정 결과

(그림 11)은 이를 바탕으로 유사도가 가장 높은 사용자 3명을 뽑아 사용자에게 결과를 추천한 결과화면이다.

(그림 12) 도서추천 결과화면

(그림 12)는 또 다른 사용자인 경기도에 사는 관심분야를 자연과학으로 설정한 18살 여성 사용자의 유사도 측정 결과이다.



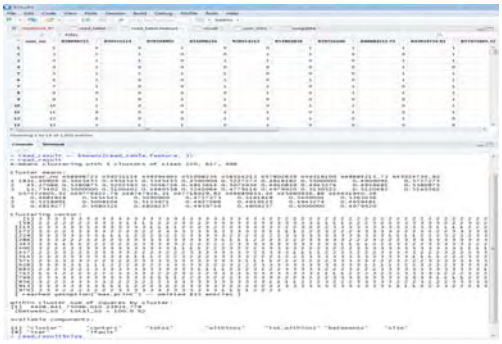
user_no	sex	age	part	location	read_book	read_book.1	read_book.2	similarity	total
1	20	2	17	2	1	26031499	58869458	59394905	0.9821	0.9934
2	42	2	19	2	2	51098234	58514213	57802839	0.9729	0.9837
3	60	2	20	3	2	59316100	59268091	45513086	0.8565	0.8642

(그림 13) 여자 사용자 유사도 측정 결과

(그림 14)은 이를 바탕으로 유사도가 가장 높은 사용자 3명을 뽑아 사용자에게 결과를 추천한 결과화면이다.

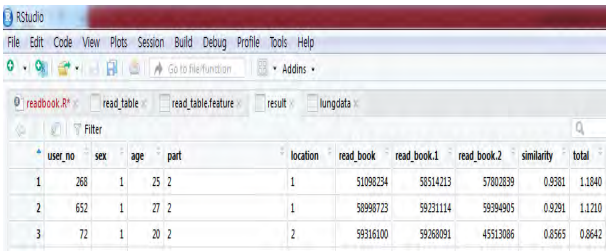
(그림 13) 도서추천 결과화면

이번에는 사용자들을 먼저 군집화 시키고 새로운 사용자에 대해서 군집들과 유사도를 계산하여 추천하는 시스템을 구현한다. k-means군집화를 하기 위해 R-studio에서 제공하는 kmeans function을 사용하여 군집화를 했다. 이를 이용하기 위해서 사용자와 사용자가 읽은 책을 이차원적으로 저장하였고 이를 기반으로 군집화를 했다. (그림 15)는 사용자가 읽은 책과 이를 기반으로 군집화를 한 모습이다.



(그림 15) 사용자 도서 기록 행렬과 군집화

군집화 후 처음 유사도를 측정했던 서울에 사는 관심분야를 문학으로 설정한 26살 남자 사용자에 대한 유사도 측정을 한 결과 (그림 16)과 같다.



	user_no	sex	age	part	location	read_book	read_book.1	read_book.2	similarity	total
1	268	1	25	2	1	51098234	58514213	57802839	0.9381	1.1840
2	652	1	27	2	1	58998723	59231114	59394905	0.9291	1.1210
3	72	1	20	2	2	59316100	59268091	45513086	0.8565	0.8642

(그림 16) 군집화 후 남자 사용자 유사도 측정

(그림 17)은 이를 바탕으로 유사도가 가장 높은 사용자 3명을 뽑아 사용자에게 결과를 추천한 결과화면이다.

(그림 17) 군집화 후 도서추천 결과화면

본 논문에서는 (그림 14)와 같이 MAE를 이용하여 정확도를 측정하였다.

$$MAE = \frac{\sum_{i=1}^{q} |\text{실제고객평가치}i - \text{예측된평가치}i|}{q}$$

(그림 14) Mean Absolute Error[8]

q는 사용자가 평가한 아이템의 개수이며, 실제 고객평가치i는 고객이 평가한 i번째 아이템의 경험 유무를 의미한다. 사용자가 I번째 아이템을 읽었다면 1, 아이템에 대한 평가가 공백이라면, 즉 읽지 않는다면 0으로 처리한다. 같은 의미로 예측된 평가치i는 사용자를 제외한 근접 이웃의 경험 유무의 평균으로서 사용자의 경험 유무를 예측한 것이다. 따라서 MAE는 실제 값과, 경험 예측치의 오차의 합을 q로 나눈 예측치 오차의 평균으로 예측한 평가치와 실제 고객의 평가치의 오차를 나타내는 지표이다 [8].

기존의 협업 필터링 방법은 사용자들간의 유사도를 측정하여 도서를 추천하였지만, 추천의 정확성을 증가시키기 위해 추가적으로 선호장르, 감정어카운팅을 통해 추천한 테스트데이터 4,000건의 평균 MAE의 비교는 다음 (표 1)과 같다.

(표 1) 평균 MAE 비교

Non-Nouncount MAE avg	Use-Nouncount MAE avg
0.3602	0.2748

이전의 평범한 협업 필터링으로 추천해준 결과보다 MAE가 0.09 줄어든 것이 보였습니다. 즉 선호장르 감정어 카운팅을 한 것이 더 신뢰성이 높음을 알수 있습니다.

4. 결론

본 논문에서는 증가하고 있는 책의 데이터로 인하여 선택하기 어려워하는 사람들을 위하여 추천 시스템을 구현했다. 또한 협업필터링의 추천정확률을 증가시키기 위하여 관심분야 단어카운팅을 통한 유사도 추천을 하여 추천 정확률을 증가시켰다.

향후 연구 과제로는 본 추천시스템의 추천정확률을 올리기 위하여 관심분야 관련 단어를 유추해내는 부분에 대해 많은 연구가 필요하다. 또한 현재 사용자 DB가 많이 부족한 상황이므로 더욱 많은 사람들이 사용하고 평가해 봐야 할 것이다.

참고문헌

[1] 임승영, “소셜 태깅 정보를 이용한 책 추천 시스템”,이화여자대학교 대학원 2018.2

[2] 정경숙, “빅데이터 기반 도서추천시스템의 설계에 관한 연구” 2017.8.

[3] http://ko.wikipedia.org/wiki/%ED%98%91%EC%97%85_%ED%95%84%ED%84%B0%EB%A7%81

[4]고경민, 모바일 환경에서 상황정보를 이용한 하이브리드 필터링 추천 시스템 설계, 한성대학교, 2010

[5] 정경숙, “빅데이터 기반 도서추천시스템의 설계에 관한 연구” 2017.8

[6] 고경민, 모바일 환경에서 상황정보를 이용한 하이브리드 필터링 추천 시스템 설계, 한성대학교, 2010

[7] 박용준. (2015). 트렌드와 고장 예측 능력을 반영한 소프트웨어 신뢰도 성장 모델 선택 방법, 한국정보과학회

[8] 심대수, R에서 협업 필터링과 개인화 요인을 이용한 개인화 영화추천 시스템, 순천향대학교, 2017